

C02029: Doctor of Philosophy
Subject Code: 32903
February 2020

CRICOS Code: 009469A

*Game theoretical adversarial deep learning
algorithms
for robust neural network models*

Aneesh Sreevallabh Chivukula

Advanced Analytics Institute, School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

Game theoretical adversarial deep learning algorithms for robust neural network models

*A thesis submitted in partial fulfilment of the requirements
for the degree of*

Doctor of Philosophy
in
Analytics

by

Aneesh Sreevallabh Chivukula

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

February 2020

ABSTRACT

Despite recent advances in deep learning, there is still a significant gap between the robustness of human perception and machine intelligence. Deep learning is not provably secure. In fact, deep neural networks are vulnerable to security attacks from malicious adversaries, which is an ongoing and critical challenge for deep learning researchers. Even innocuous perturbations in training data can change the way a deep network behaves in unintended ways. This means that imperceptibly and immeasurably small departures from the training data can result in a completely different label classification when using the model for supervised deep learning.

In this thesis, we explore adversarial deep learning algorithms. We examine how they exploit vulnerabilities in deep networks and how to make deep networks robust to their attacks. To explore the vulnerabilities, we simulate various model training processes under a range of various attack scenarios. Each attack strategy is assumed to be formulated by an intelligent adversary that is capable of either feature manipulation, label manipulation, or both. The optimal attack policy of our adversaries is determined by the solution for optimization problems that output the adversarial data. We then apply the knowledge that we learned to improve and reinforce the learning procedure so as to better defend against attacks.

As part of this research process, we developed new adversarial learning algorithms to solve for adversarial manipulations in supervised classification networks such as Convolutional Neural Networks (CNNs). The adversarial learning objective for our adversaries is to inject small changes into the data distributions, defined over positive and negative class labels, to the extent that the CNN subsequently misclassifies the data distribution. Thus, the theoretical goal of our deep learning process becomes one of determining whether a manipulation of the input data has reached a learner decision boundary, i.e., where too many positive labels have become negative labels. We began this research undertaking by first studying the performance vulnerabilities in CNNs. With these vulnerabilities identified, we were able to propose CNNs that are secure to those types of adversarial attacks.

We generate adversarial data by solving for optimal attack policies in Stackelberg games where adversaries target the misclassification performance of CNNs. In a sequential game-theoretic formulation, we model the interaction between an intelligent adversary and a deep learning model (a CNN) to generate adversarial manipulations by solving a two-player sequential noncooperative Stackelberg game where each player's payoff function increases with interactions to a local optimum. With a stochastic game-

theoretic formulation, we then extend the two-player Stackelberg game into a multiplayer Stackelberg game with stochastic payoff functions for the adversaries. Both versions of the game are resolved through the Nash equilibrium, which refers to a pair of strategies in which there is no incentive for either the learner or the adversary to deviate from their optimal strategy. In this case, the strategy pair is a learner weight and an evolutionary operation. In addition, we devised each attack scenario as a blackbox attack where the adversaries have no prior knowledge of the CNN’s learning processes and its best response strategies. In each case, we show that the Nash equilibrium leads to a supervised classification network that is robust to subsequent data manipulation by a game-theoretic adversary.

We then explore adversaries who optimise variational payoff functions via data randomisation strategies on CNNs designed for multilabel classification tasks. Similarly, the outcome of these investigations is an algorithm design that solves a variable-sum two-player sequential Stackelberg game with new Nash equilibria. In designing the attack scenarios, the adversarial objective was to make small, undetectable changes to the test data. The adversary manipulates variational parameters in the input data to mislead the learning process of the CNN, so it misclassifies the original class labels as the targeted class labels. Our ideal variational adversarial manipulation is the minimum change needed to the adversarial cost function of encoded data that will result in the CNN incorrectly labelling the decoded data.

The resulting attack algorithms were a Simulated Annealing (SA) algorithm and an Alternating Least Squares (ALS) algorithm, which optimise the data manipulations by stochastically searching the strategy spaces of the variational adversaries. The optimal manipulations are found by solving the Nash equilibria optimization problem of the proposed Stackelberg game. Specifically, the optimal attack parameters in ALS and SA solve for the (variational nonlinear non-convex) adversarial cost functions that generate adversarial data. The payoff function for the variational adversary depends on the manipulations determined by a Variational Autoencoder (VAE), while the payoff function for the CNN classifier is evaluated in the input data space.

The adversarial data generated by this variant of the Stackelberg games simulates continuous interactions with the classifier’s learning processes as opposed to one-time interactions. To evaluate CNN performance, we assessed the adversarial algorithms over different strategy spaces proposed for the MNIST handwritten digits database and the VGGFace2 database of human faces.

The original CNN model was then retrained using all the adversarial manipulations generated by the game-theoretic adversaries to create a secure CNN model. In an empirical demonstration, we show that the new secure CNN model is robust to subsequent game-theoretic data manipulation by adversaries. This promising result suggests that evolutionary algorithms based on game-theoretic modelling and mathematical optimization are significantly better approach to building more secure deep learning models.

We also empirically demonstrate that variational adversaries are also able to mislead CNNs. In these cases, the learning process of the CNNs was manipulated by an adversary at the input data level as well as the generated data. The optimal manipulations were

to stochastic optima in non-convex best responses strategies. We were able to encode the resulting adversarial data in terms of the multivariate statistical parameters of a Gaussian mixture model. We then retrained the original CNN on the manipulated data to give rise to a secure CNN that is robust to subsequent performance vulnerabilities from variational adversaries.

In applying this research, we developed a deep network model that discovers Granger causes in multivariate temporal financial market data. The model comprises a deep neural network (DNN) and a recurrent neural network (RNN) and discovers Granger-causal features with bivariate regression from bivariate time series data distributions. These features are subsequently used to discover Granger-causal graphs for multivariate regression on multivariate time series data distributions. Our supervised feature learning process with these proposed deep regression networks returned favourable F-tests for feature selection and t-tests against comparisons with other models. Moreover, a regression analysis on a set of experiments with real stock market data from Yahoo Finance demonstrates that our causal features are a significant improvement over the existing deep learning regression models in terms of minimising root mean squared error.

CERTIFICATE OF AUTHORSHIP

I, *Aneesh Chivukula* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science, Faculty of Engineering & Information Technology* at the University of Technology Sydney, Australia, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

[Aneesh Sreevallabh Chivukula]

DATE: 10th February, 2020

PLACE: Sydney, Australia

DEDICATION

To my parents . . .

कारणाभावात् कार्याभावः ।

The absence of the effect is due to the absence of the cause.

Vaisesika Sutra 1.2.1 enunciating the Causality Principle in Indian Philosophy.

ACKNOWLEDGMENTS

I want to thank my principal supervisor, Dr. Wei Liu for his strong support and guidance throughout my degree. His scientific expertise and continuous encouragement of my research have been essential for my achievements. I also want to thank my co-supervisor, Dr. Jun Li who organized a friendly research environment and supported my casual academics at UTS.

I would like to thank Capital Markets CRC Limited (now RoZetta Institute) for providing me with a PhD scholarship. I would like to thank Australian Mathematical Sciences Institute for providing me with a postgraduate research internship.

Finally, I want to thank my wife for her support throughout my PhD.

I acknowledge Chandranath Adak (UTS) for providing this thesis template.

LIST OF PUBLICATIONS

RELATED TO THE THESIS :

1. A. S. Chivukula and W. Liu, "Adversarial learning games with deep learning models," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 2758-2767. doi: 10.1109/IJCNN.2017.7966196
2. A. S. Chivukula and W. Liu, "Adversarial Deep Learning Models with Multiple Adversaries," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 6, pp. 1066-1079, 1 June 2019. doi: 10.1109/TKDE.2018.2851247
3. Chivukula A.S., Li J., Liu W. (2018) Discovering Granger-Causal Features from Deep Learning Networks. In: Mitrovic T., Xue B., Li X. (eds) AI 2018: Advances in Artificial Intelligence. AI 2018. Lecture Notes in Computer Science, vol 11320. Springer, Cham doi: 10.1007/978-3-030-03991-2_62
4. A. S. Chivukula, X. Yang and W. Liu, "Adversarial Deep Learning with Stackelberg Games," accepted for presentation at the 26th International Conference on Neural Information Processing (ICONIP 2019) of the Asia-Pacific Neural Network Society.
5. A. S. Chivukula, X. Yang and W. Liu, "Game Theoretical Adversarial Deep Learning with Variational Adversaries," in IEEE Transactions on Knowledge and Data Engineering (TKDE), doi: 10.1109/TKDE.2020.2972320.

OTHERS :

6. Prabhu, C.S.R., Sreevallabh Chivukula, A., Mogadala, A., Ghosh, R., Livingston, L.M.J. 2019. Big Data Analytics: Systems, Algorithms, Applications. Springer Nature Singapore Pte Ltd.

TABLE OF CONTENTS

List of Publications	xi
List of Figures	xvii
List of Tables	xix
1 Introduction	1
1.1 Thesis and Contributions	2
1.2 Research Methods, Objectives and Significance	6
1.3 Report Organisation	8
I	9
2 Related Work	11
2.1 Adversarial Machine Learning	12
2.1.1 Adversarial Security Mechanisms	16
2.1.2 Adversarial Learning Frameworks	18
2.2 Adversarial Deep Learning	21
2.2.1 Adversarial Examples in Deep Networks	22
2.2.2 Adversarial Examples for Misleading Deep Classifiers	24
2.2.3 Generative Adversarial Networks	27
2.2.4 Generative Adversarial Networks for Adversarial Learning	28
2.2.5 Causal Inference in Deep Learning	34
2.2.6 Causal Inference in Time Series Analysis	34
2.2.7 Causal Inference in Financial Markets	35
2.2.8 Causal Feature Learning and Adversarial Machine Learning	36
2.2.9 Explainable Artificial Intelligence and Adversarial Machine Learning	37
2.3 Spam Filtering	38

TABLE OF CONTENTS

2.3.1	Biometric Spam	38
2.3.2	Image Spam	39
2.3.3	Text Spam	40
2.4	Security and Privacy in Adversarial Learning	40
2.4.1	Adversarial Attack Scenarios in Linear Classifiers	42
2.4.2	Adversarial Attack Scenarios in Feature Weighting	43
2.4.3	Poisoning Support Vector Machines	45
2.4.4	Robust Classifier Ensembles	47
2.4.5	Robust Clustering Models	48
2.4.6	Robust Feature Selection Models	49
2.4.7	Robust Anomaly Detection Models	50
2.4.8	Robust Task Relationships Models	52
2.4.9	Adversarial Machine Learning in Cybersecurity	53
2.5	Game Theoretical Adversarial Deep Learning Models	54
2.5.1	Game-Theoretic Learning Models	54
2.5.2	Game theoretical adversarial learning	55
2.5.3	Game theoretical adversarial deep learning	57
2.5.4	Stochastic Games in Predictive modelling	59
2.5.5	Nash equilibria and Stackelberg Strategies in Differential Games	61
2.6	Adversarial Defense mechanisms	62
2.6.1	Securing classifiers against feature attacks	62
2.6.2	Adversarial classification tasks with regularizers	64
2.6.3	Adversarial Reinforcement Learning	64
2.7	Computational optimization algorithms	65
2.7.1	Derivative-free Stochastic Optimization	67
3	Game Theoretical Adversarial Deep Learning with Evolutionary Adversaries and Stochastic Adversaries	69
3.1	Sequential game formulation	69
3.2	Stochastic game formulation	72
3.3	Stochastic game illustration	73
3.3.1	Illustrative Adversarial Examples	75
3.4	Stochastic Game Algorithm	76
3.5	Sequential Game Algorithm	77
3.5.1	Genetic algorithm	77

3.5.2	Simulated annealing algorithm	80
3.6	Experiments	81
3.6.1	Dataset description	83
3.6.2	Genetic algorithm validation in Sequential Game	83
3.6.3	Simulated annealing validation in Sequential Game	87
3.6.4	Fitness function validation in Sequential Game	88
3.6.5	Evolutionary operations validation in Stochastic Game	92
3.7	Findings Summary	94
4	Game Theoretical Adversarial Deep Learning with Randomization Strategies	97
4.1	The differences between adversarial data generated by randomization strategies and evolutionary strategies	97
4.2	Game Formulation	98
4.2.1	Stackelberg game formulation	98
4.2.2	Stackelberg Game Illustration	101
4.3	Our Proposed Algorithms	102
4.3.1	Adversarial Learning Algorithm	103
4.3.2	Adversarial Manipulations Generation	103
4.3.3	Simulated Annealing Algorithm	104
4.4	Experiments	105
4.4.1	Classifier and Autoencoder Description	105
4.4.2	Attack and Defence Performance Validation with Fixed Target and Original Data	105
4.4.3	Defence Performance Validation with Varying Target and Generated Data	108
4.5	Findings Summary	109
5	Game Theoretical Adversarial Deep Learning with Variational Adversaries	111
5.1	Game Formulation	111
5.1.1	Overall Structure of Our Model	117
5.1.2	The differences between our method and GANs	118
5.1.3	Variational Game and Adversarial Examples Illustration	119
5.2	Variational Stackelberg Game Method	119
5.2.1	Alternating Least Squares Algorithm	123

TABLE OF CONTENTS

5.2.2	Simulated Annealing Algorithm	124
5.3	Experiments	125
5.3.1	Classifier and Autoencoder Description	125
5.3.2	Attack Performance Validation	128
5.3.3	Defence Performance Validation	129
5.4	Findings Summary	132
II		135
6	Discovering Granger-causal Features with Deep Learning Networks	137
6.1	Our Proposed Algorithms	138
6.1.1	Empirical Risk training in Deep Learning Networks	138
6.1.2	Granger Causality testing in Deep Learning Networks	139
6.1.3	Multivariate Regression validation with Deep Learning Networks	140
6.1.4	Deep Learning Networks based Regression Models	141
6.2	Experiments	144
6.2.1	Single Granger-causes validation	145
6.2.2	Multiple Granger-causes validation	147
6.3	Findings Summary	148
7	Conclusion and Future Work	151
7.1	Research Summary	151
7.2	Applications of Game Theoretical Adversarial Learning Models	155
	Bibliography	161

LIST OF FIGURES

FIGURE	Page
3.1 A flow chart illustrating the benefits of a game theoretic learner. The two-player game is played by a single adversary and one Learner. The game produces a final deep learning network CNN_{secure} that is better equipped to deal with the adversarial manipulations than the initial deep learning network $CNN_{original}$	74
3.2 Examples of transformed images found at Nash equilibrium in a Stackelberg game. To avoid detection, the adversary adds pixels in (a) and (d), and changes shape in (b) and (c).	75
3.3 Examples of transformed images found at Nash equilibrium in a Stackelberg game. To avoid detection, the adversary adds pixels in (a), changes shape in (b), deletes pixels and changes shape in (c) and (d)	76
3.4 Testing performance with variations in evolutionary operators consisting of genetic parameters and annealing parameters. Genetic parameters are given in fig(a), fig(b), fig(c), and fig(d) for mutation step δ , crossover width η , selection size ζ , and population size ψ respectively. Annealing parameters are given in fig(e), fig(f), fig(g), and fig(h) for annealing step δ , annealing mask width η , annealing sample size ν , and annealing reduction rate ρ respectively. The manipulated learner has lower performance than the original learner. The secure learner has higher performance than the manipulated learner.	84
3.4 Testing performance with variations in evolutionary operators consisting of genetic parameters and annealing parameters. Genetic parameters are given in fig(a), fig(b), fig(c), and fig(d) for mutation step δ , crossover width η , selection size ζ , and population size ψ respectively. Annealing parameters are given in fig(e), fig(f), fig(g), and fig(h) for annealing step δ , annealing mask width η , annealing sample size ν , and annealing reduction rate ρ respectively. The manipulated learner has lower performance than the original learner. The secure learner has higher performance than the manipulated learner.	85

3.5	Testing performance with variation in error weight λ . The manipulated learner has lower performance than the original learner. The secure learner has higher performance than the manipulated learner.	89
4.1	A flowchart illustrating the adversarial autoencoder based Stackelberg game-theoretic modelling.	101
4.2	Testing performance (error) with variations in attack operators consisting of adversarial cost weight λ and autoencoder code size ρ	106
4.2	Testing performance (error) with variations in attack operators consisting of adversarial cost weight λ and autoencoder code size ρ	107
5.1	A flowchart illustrating the variational Stackelberg game theoretical adversarial learning.	117
5.2	Examples of transformed images found at Nash equilibrium in a Stackelberg game. For the two images in each subfigure, the left one is the original image and the right is the manipulated image.	120
5.3	MNIST Testing performance (error) with variations in attack parameters consisting of encoder code size s , adversarial cost weight λ and annealing steps limit N	126
5.4	VGGFace2 Testing performance (error) with variations in attack parameters consisting of encoder code size s , adversarial cost weight λ and annealing steps limit N	127
6.1	Granger-causal features, F-statistics on RMSEs $RMSE_r, RMSE_{ur}$ and multivariate regression RMSEs $RMSE_{mv}$ for the unrestricted model with DNN. The edge directions indicate the causal relations between pairs of stocks and the edge weights show the corresponding F-test statistic given in Definition 7.	147

LIST OF TABLES

TABLE	Page
2.1 Adversarial Algorithms Comparison 1	14
2.2 Adversarial Algorithms Comparison 2	15
2.3 Generative Adversarial Networks Comparison 1	30
2.4 Generative Adversarial Networks Comparison 2	31
2.5 Generative Adversarial Networks Comparison 3	32
2.6 Generative Adversarial Networks Comparison 4	33
3.1 Datasets of colour images used in the experiments	83
3.2 Genetic Algorithm : p-values comparison before and after game by varying parameters for each genetic operator. t-statistics are computed between pairs of learner F1-score performance, manipulated learner F1-score performance and secure learner F1-score performance. CNN_o, CNN_m and CNN_s are short names for $CNN_{original}, CNN_{manipulated}$ and CNN_{secure} respectively.	91
3.3 Simulated Annealing Algorithm : p-values comparison before and after game by varying parameters for each annealing operator. t-statistics are computed between pairs of learner F1-score performance, manipulated learner F1-score performance and secure learner F1-score performance. CNN_o, CNN_m and CNN_s are short names for $CNN_{original}, CNN_{manipulated}$ and CNN_{secure} respectively.	91
3.4 Two-player Two-Label Sequential Games Performance Evaluation - F1-score before and after two-player game across various combinations of handwritten digits. CNN_o, CNN_m and CNN_s are short names for $CNN_{original}, CNN_{manipulated}$ and CNN_{secure} respectively. We observe a consistent decrease in the manipulated learner CNN_m performance tested on adversarial data as compared to learner CNN_o performance. We also observe a consistent increase in the secure learner CNN_s performance compared to manipulated learner CNN_m	93

3.5	Multiplayer Two-label Stochastic Games Performance Evaluation - F1-score before and after multiplayer game across various combinations of handwritten digits. CNN_o, CNN_m and CNN_s are short names for $CNN_{original}, CNN_{manipulated}$ and CNN_{secure} respectively. We observe a consistent decrease in the manipulated learner CNN_m performance tested on adversarial data as compared to learner CNN_o performance. We also observe a consistent increase in the secure learner CNN_s performance compared to manipulated learner CNN_m	94
4.1	Autoencoder Attack Scenario: t-tests before and after game by varying parameters for target class “7”.	108
4.2	Comparisons on the defence to adversarial Nash equilibrium attacks	109
5.1	Table of Notation	113
5.2	Alternating Least Squares Attack Scenario: t-tests by varying parameters in Variational Stackelberg games. The small p -values from the statistical tests demonstrate the superiority of our model.	129
5.3	MNIST Comparisons on the defence to adversarial Nash equilibrium attacks. . . .	129
5.4	VGGFace2 Comparisons on the defence to adversarial Nash equilibrium attacks . .	129
6.1	Companies Listing	144
6.2	RMSEs with MSE loss for bivariate regression. DNN is selected as the best network structure for Granger causality.	145
6.3	RMSEs with MSE loss for Granger-causal feature discovery. The rows show causal relations with the restricted model and the unrestricted model RMSEs $RMSE_r$ and $RMSE_{ur}$ in bivariate regression with DNN.	146
6.4	RMSEs with MSE loss for Granger-causal feature discovery. The rows show causal relations with the restricted model and the unrestricted model RMSEs $RMSE_r$ and $RMSE_{ur}$ in bivariate regression with RNN.	146